DOE National Collaboratories Meeting, August 10-12, Argonne National Laboratory

A National Science Grid in Support of Future Facilities: An Urgent Need for DOE

Ian Foster, foster@mcs.anl.gov

DOE Office of Science laboratories operate a wide range of unique resources, from light sources to supercomputers and petabyte storage systems, that are used by large user communities. The laboratories' geographically distributed staff are frequently faced with scientific and engineering problems of great complexity, the solution of which requires the creation and effective operation of large multidisciplinary teams. These teams must address large and challenging problems that often exceed the limits of traditional computing and information systems approaches.

These three defining characteristics of DOE laboratories and DOE programs—unique resources established for remote users, large distributed teams, and ultra-large-scale computing and data—will only be accentuated as the ambitious facilities described in the Office of Science's 20 year plan come online. Thus it is timely to ask: will we embed these revolutionary 21st Century facilities within a 21st Century science infrastructure, or will we continue to rely on 19th Century mechanisms (personal visits, transport of data by mail) as our primary means of communication? Will fusion scientists participate in the operation of ITER experiments as first class citizens, within a virtual control room, or will they be forced travel to a remote laboratory? Will computational biologists be able to analyze all of the data produced by their simulations, as it is produced, or be forced to wait for a subset to arrive by federal express? Will users of facilities such as the SNS detect and correct configuration errors while running experiments, or realize days after an experiment has occurred that their data is worthless?

The answer to these questions depends on whether we make commitments now to invest in the design, development, deployment, and operation of a 21st Century distributed computing infrastructure that will eliminate barriers to remote resource access and enable the coupling and integration of geographically distributed resources and expertise. The facilities roadmap outlines one critical infrastructure requirement for future facilities, namely the rapid expansion of ESnet towards ultra-high-speed operation. However, we need far more than fast networks: we need a *National Science Grid* that will support a dramatically enhanced coupling of the resources and scientists of DOE laboratories with each other and with the national research community. This Grid will enable virtual control rooms for international experiments, real-time analysis of data from computational experiments, remote participation in remote experiments, and collaborative data analysis within distributed teams.

The "DOE Science Grid" project funded under the National Collaboratories program has made useful steps towards the realization of this overall goal, via the development, deployment, and operation of relevant system components, notably an authentication service, as well as outreach to key application communities. Other projects, notably the Particle Physics Data Grid and its Trillium partners GriPhyN and iVDGL, have taken significant steps towards the realization of a distributed infrastructure for data-intensive science (e.g., see the Grid3 system), while the Earth System Grid, Fusion Collaboratory, and the Collaboratory for Multi-scale Chemical Science are all offering production services to their communities.

The DOE National Collaboratories program should be proud of these and other related successes, which have been instrumental in transforming the notion of a "national grid" or "cyberinfrastructure" from outlandish vision to practical concept. However, the translation of this concept to production infrastructure now requires substantial further effort and financial support, if we are to realize the benefits reviewed above. I point out five key challenges in the following.

1) *Grid-enabled science*. Despite the successes of DOE and other collaboratory projects, it is still the case that the vast majority of DOE science are not exploiting collaboratory techniques to the extent that they could. Further penetration of the DOE science community will require not only further application partnerships but also further research aimed at understanding the work processes of different communities and the ways in which technology can further those work processes.

2) *Security and trust*. The network environment in which collaboratory applications operate is becoming more dangerous as attacks become more sophisticated. Concurrently, larger and more diverse user communities and more challenging resource sharing modalities and creating new demands for trust and security mechanisms. Major advances are going to be required in the practice and probably also the theory of security, trust management, intrusion detection, and related fields. The demands of collaborative science overlap with, but are not the same as, those encountered in industry. DOE must invest in this area if the required progress is to occur.

3) *Knowledge-driven collaboration*. Collaborative science is ultimately about the creation, sharing, and evolution of knowledge artifacts such as data and programs. Yet the tools we have for creating, annotating, discovering, transforming, explaining, archiving, etc., such artifacts are incredibly primitive, particularly within the context of multi-institutional virtual organizations.

4) *Deployment and operations*. Just as networks were first deployed by researchers and then transitioned to site infrastructure groups, so Grid facilities and services need to become the responsibility of laboratory and DOE-wide infrastructure operators. This transition will require not only human resources and hardware, but also innovation in areas of operational procedures and policies, so that key functions such as security can be handled effectively in a distributed setting.

5) *Sensors*. DOE laboratories and scientists have not yet made significant use of recent advances in sensor and wireless networking technology. Yet these technologies seem likely to revolutionize many aspects of DOE science, certainly in environmental studies and presumably in other domains as well.

The realization of these goals will require significant investments in three areas, each of critical and equal importance to the overall endeavor: infrastructure development, applications partnerships, and technology research. *Infrastructure development* will put in place persistent services such as authentication, resource discovery, resource management, and remote data access. This work is important because without it, individual tool and application efforts are forced to reinvent crucial services. *Application partnerships* will bring together application and distributed computing scientists to pioneer new approaches to scientific computing and to introduce Grid computing to new communities. Finally, *technology research* will contribute to the development of the tools, services, algorithms, and concepts required to create next-generation Grid infrastructures and enable new classes of Grid applications.